Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Bayesian Modelling of Nonstationary Extreme Values Using a Nonparametric Hawkes Process

Gordon J. Ross

gordon.ross@ed.ac.uk

University of Edinburgh

Dean Markwick

University College London

Research Article

Keywords: Hawkes Processes, Extreme Values, MCMC

Posted Date: May 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9596512/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Statistics and Computing on October 6th, 2026. See the published version at https://doi.org/10.1007/s11222-026-10981-y.

Bayesian Modelling of Nonstationary Extreme Values Using a Nonparametric Hawkes Process

Gordon J. Ross[1*] and Dean Markwick[2]

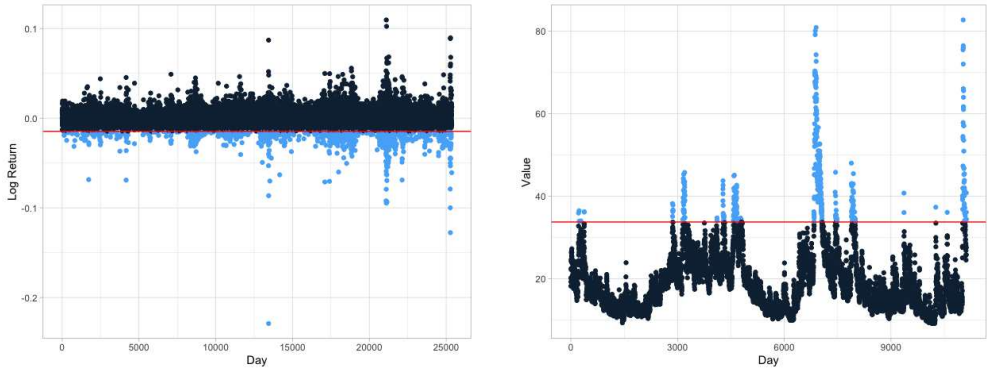
[1*]Department of Mathematics, University of Edinburgh, UK.
[2]Department of Statistics, University College London, UK.

*Corresponding author(s). E-mail(s): gordon.ross@ed.ac.uk;
Contributing authors: dean.markwick@talk21.com;

Abstract

Modelling and forecasting the occurrence of extreme events is especially difficult when the event process is nonstationary, with changes in both the rate at which extremes occur and the magnitude of the extremes when they occur. We approach this task by developing a Bayesian point process model for extreme events, which uses a self-exciting Hawkes process to model the rate at which extremes occur. The Hawkes process has a structure which allows events to occur in clusters, making it realistic for many types of data. We use a flexible Bayesian nonparametric approach based on the Dirichlet process to learn the temporal excitation pattern from the data. Further, we build on Extreme Value Theory by using a Generalised Pareto Distribution (GPD) to model the magnitudes of the extremes, with a hierarchical mark model allowing these magnitudes to vary across Hawkes-induced clusters. A hierarchical specification of the model results in partial pooling, allowing for more accurate GPD estimation even in clusters with only a small number of observations. We develop an MCMC algorithm to sample from the resulting hierarchical model. A simulation study confirms that the two flexible components improve prediction when the corresponding features are present in the data-generating mechanism, and across four real data sets the nonparametric Hawkes model with hierarchical GPD marks gives the best held-out predictive performance among the model variants considered.

Keywords: Hawkes Processes, Extreme Values, MCMC



(a) Daily log returns of the S&P 500 index from 1951–2020.

(b) Daily closing values of the VIX index from 1990–2020.

Fig. 1: Illustration of two stock market indexes with time variation in both the rate at which extreme values occur, and their magnitudes when they do occur. Extremes are defined as values above the 95^{th} percentile of the distribution (negative for the $S\&P$ 500) and colored in light blue.

1 Introduction

Effective risk management often requires an estimate of the probability that large events will occur during a given period of time. For example, suppose that m terrorist attacks have previously occurred over a period of T years, at times $t_1, t_2, \ldots, t_m$. For each attack at time t_i, let r_i be a mark denoting the corresponding number of fatalities. Based on this historical data, it may be desirable to produce a probabilistic estimate for the probability of another large attack occurring within some future time window (Porter and White 2012). Similar problems are also often considered in fields such as natural hazards modelling where the events correspond to earthquakes and the marks correspond to earthquake magnitudes (Bray and Schoenberg 2013), and in finance where the events are the times at which large changes in a company's stock price are observed, and the marks represent the size of the change (Kiriliouk et al. 2019; Chavez-Demoulin and McGill 2012).

This prediction task can often be broken down into two parts. First, a prediction is made for the number of large events which are likely to occur in a given period along with their occurrence times. Second, for each predicted event r_i, the probability $p(r_i > z)$ that the mark exceeds a threshold z is then estimated. Direct estimation of this latter quantity is difficult since it usually involves inference about extreme quantiles of the distribution governing r_i. This can be highly sensitive to the parametric assumptions made about the distribution, with even small misspecification giving very inaccurate results. As such, it is usual to instead rely on asymptotic results from the field of Extreme Value Theory (EVT) to avoid the need to make strong parametric assumptions (Coles 2001; Davison and Smith 1990).

For this purpose, the well-known Pickands–Balkema–de Haan (PBH) theorem from EVT states that given a sufficiently large threshold u, the conditional excess distribution $p(r_i - u \leq z \mid r_i > u)$ can be approximated by a Generalised Pareto Distribution (GPD), as long as the distribution satisfies certain regularity conditions (Balkema and de Haan 1974; Davison and Smith 1990). This has led to the widely-used Peaks-Over-Threshold (POT) approach for modelling extreme values, where the occurrence times t_i of extremes are commonly modelled using a Poisson process, with the corresponding excess magnitudes $r_i - u$ following a GPD (Chavez-Demoulin et al. 2005). However the direct application of this methodology to real-world problems is limited by the strong stationarity assumptions that it requires (Coles 2001). Specifically, a straightforward EVT analysis is only possible if: a) the occurrence times t_i of large values can be modelled by a possibly inhomogeneous Poisson process and b) the magnitudes exceeding the threshold are independent draws from a GPD with constant parameters.

Unfortunately these assumptions often do not hold. Figure 1 shows the daily values of two major stock market indexes. Both indexes are designed to give an easily calculable measure of the US stock market and an indicator of the general outlook and volatility of the economy. Many risk measures are based on the S&P 500 daily closing price, and it is therefore of interest to predict when the next large drop could occur and how severe it could be (Chavez-Demoulin et al. 2005). In Figure 1 we can see notable clustering of the extremes, where they tend to occur in bursts. Furthermore, there appears to be structural change in the distribution $p(r_i > z)$ of the extremes, with losses being structurally larger during certain time periods as well as more frequent. This is partly due to the well-known phenomenon of heteroskedasticity in stock returns as discussed by Poon et al. (2003). It would hence be unwise to fit a single stationary GPD to these data.

Numerous extensions of EVT have been proposed to model non-stationarity in extreme events. An influential early paper suggested a parametric regression framework for the GPD parameters to allow variation over time (Davison and Smith 1990) and this idea has been extended in several ways (Northrop and Jonathan 2011; Chavez-Demoulin et al. 2005). Related work has proposed various more sophisticated models for the point process governing the occurrence of extremes, for example (Gyarmati-Szabo 2011; Stindl 2023), while other approaches involve an initial pre-processing step aimed at removing nonstationarity (Eastoe and Tawn 2009). A partial review of the extensive literature on EVT can be found in Coles (2001).

Building on the marked point process representation of threshold exceedances, Kottas and Sansó (2007) and Kottas et al. (2012) developed Bayesian nonparametric models for the associated intensity function. The flexibility of nonparametric estimation allows nonstationarity in both the time and mark domains to be handled naturally. However while their framework is well-suited to modelling historical data, it is less useful for making predictions about the occurrence of extremes in the future. This is because their point process representation effectively smooths out the historical data rather than explicitly modelling the conditional intensity function of the point process, which makes it difficult to make predictions based on recent process behaviour.

This article develops a Bayesian marked Hawkes POT model for predicting future extreme events in situations where the exceedances are non-stationary in both the

time and mark domains. The occurrence times of exceedances are modelled using a self-exciting Hawkes process (Hawkes 1971), so that recent exceedances can increase the short-term probability of further extremes and induce clusters of events. Rather than imposing a fixed parametric form on the excitation mechanism, we estimate the Hawkes triggering kernel using a flexible Bayesian nonparametric model based on the Dirichlet process. We use the latent branching representation of the Hawkes process not only for posterior computation, but also as a way of linking temporal clustering with variation in the mark distribution. Specifically, exceedances belonging to different Hawkes-induced clusters are allowed to have different mark distributions, with a hierarchical prior used to borrow strength across clusters. This links temporal clustering in the exceedance process to variation in the magnitudes of the exceedances, while avoiding the need to specify a parametric time-varying model for the GPD parameters.

The main contributions of the paper are threefold. First, we develop a marked Hawkes POT model for forecasting future threshold exceedances, rather than only smoothing historical extremes. Second, we use a Dirichlet process mixture prior for the Hawkes triggering kernel, allowing the temporal excitation pattern to be learned flexibly from the data. Third, we link the latent branching structure of the Hawkes process to a hierarchical GPD mark model, allowing the scale of exceedance magnitudes to vary across Hawkes-induced regimes while sharing information across clusters. The resulting model is evaluated using held-out predictive scores for both exceedance times and magnitudes.

We begin Section 2 by reviewing the traditional methods of EVT for estimating $p(r > z)$ when the event process is stationary. We then continue in Section 3 by introducing the Hawkes process and show how it can be used in a nonparametric manner which is suitable for when there is no strong theoretical motivation for particular parametric assumptions. In Section 4 we detail a Metropolis–Hastings algorithm for sampling the full posterior distribution of our model parameters. Section 5 evaluates the behaviour of the method in a controlled simulation study, before Section 6 applies the methodology to real-world data sets.

2 Extreme Value Theory

We first recall the standard peaks-over-threshold construction in the idealised stationary case. Suppose that $r_1, \ldots, r_m \sim F$ are a sequence of independent and identically distributed observations, and that interest lies in the probability of large values occurring. If the functional form of F is known, then this can be computed directly from its quantile function after any unknown parameters have been estimated. However the functional form of F is usually unknown, and a particular parametric form will have to be chosen based on both the observed data and theoretical considerations. Unfortunately, inference for extreme quantiles of F is known to be highly sensitive to these parametric assumptions (Porter and White 2012).

To avoid specifying a parametric model for the full distribution F, it is common to instead use the peaks-over-threshold (POT) approach, which models only the distribution of the excess $r_i - u$ conditional on $r_i > u$, where u is a threshold parameter

(Coles 2001). This approach is justified by the Pickands–Balkema–de Haan theorem (Balkema and de Haan 1974), which states that for a sufficiently large threshold u, the conditional excess distribution can be approximated by a Generalised Pareto Distribution:

Pickands–Balkema–de Haan (PBH) Theorem: Suppose $r_1, \ldots, r_m$ are i.i.d. with distribution F. Let $F_u(z) = p(r_i - u \leq z \mid r_i > u)$ denote the conditional excess distribution function which describes the behaviour of F above a given threshold u. Then, assuming that F satisfies suitable regularity conditions, F_u converges to the Generalised Pareto Distribution (GPD), i.e. $F_u(z) \to G(z \mid \sigma, \xi)$ as $u \to \infty$, where

$$G(z \mid \sigma, \xi) = \begin{cases} 1 - (1 + \xi z/\sigma)_+^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-z/\sigma), & \xi = 0, \end{cases}$$

with $z > 0$, scale parameter $\sigma > 0$, shape parameter ξ, and $(a)_+ = \max(a, 0)$. Assuming these regularity conditions are satisfied, the POT approach to extreme value estimation is to choose a threshold u sufficiently large to make the GPD a good approximation above u, estimate the GPD parameters (σ, ξ), and then approximate $p(r_i > z \mid r_i > u)$ for $z > u$ by $1 - G(z - u \mid \sigma, \xi)$.

The POT construction also has a natural marked point process interpretation (Coles 2001). Write the original data as ordered pairs (t_i, r_i) for $i = 1, 2, \ldots, m$, where t_i denotes the time at which r_i is observed. After deleting the pairs where $r_i < u$, suppose that n exceedances remain, and write these as (t_i, y_i) for $i = 1, \ldots, n$, where $y_i = r_i - u$ denotes the excess over the threshold. In the simplest stationary formulation, the exceedance times $t_1, \ldots, t_n$ are governed by a homogeneous Poisson process, while the excess magnitudes y_i are independent draws from a common GPD(σ, ξ) distribution. Equivalently, the pairs $\{(t_i, y_i)\}$ can be viewed as observations from a homogeneous marked point process whose time component is Poisson and whose mark distribution is GPD.

In this simplest formulation, the process is stationary in time: exceedances are equally likely to occur at any point in the observation window, and the excess magnitudes are independent draws from a common GPD. The model developed below relaxes both parts of this classical marked point process representation: the homogeneous Poisson process for exceedance times is replaced by a Hawkes process, and the common GPD mark distribution is replaced by a hierarchical model for clustered exceedance magnitudes.

3 Nonstationarity of the Exceedance Process

As above, let $y_1, \ldots, y_n$ denote the values of the marks which exceed some threshold u. A direct application of either the PBH theorem or the above point process representation for predicting future extreme values requires these y_i marks to be treated as independent and identically distributed, with their time occurrence following a homogeneous Poisson process. This is problematic for two reasons:

1. The point process governing the times at which the exceedances occur can be nonstationary. This was previously seen in Figure 1 where the exceedances fall into

clusters, with no exceedances occurring for long periods of time followed by many occurring close together.
2. The distribution $p(y_t)$ of the exceedances may also change over time. Again this was seen in Figure 1 where the magnitude of the extremes tends to be higher in some time periods compared to others.

A substantial amount of existing literature on extreme value theory relaxes these assumptions in various ways. For example in (Leadbetter 1976), it is shown that as long as exceedances satisfy some mixing conditions, they are no longer required to be independent and instead can display some local dependence. This local dependence is expressed through the extremal index θ (Hsing et al. 1988) and can be interpreted as the average clustering effect in the extreme values. Similarly, much of the existing literature relies on specifying parametric models for the time-evolution of both the occurrence time process, and the distribution $p(y_t)$. Although this is a reasonable approach for modelling historical non-stationarity, it typically does not allow for inference of the **conditional** exceedance distribution $p(y_t > z \mid y_t > u, y_{1:(t-1)})$ which will often be the main object of interest when making predictions about the future. In many applications it will be important to assess the probability of an extreme value occurring at some particular time point t (e.g. "next week") which requires taking into account the recent history of the process.

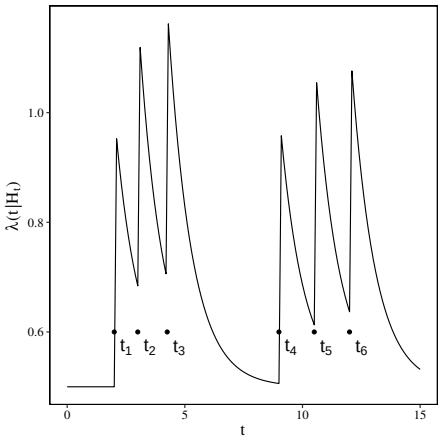
We hence take a different approach which focuses on modelling the conditional exceedance process directly. This is based on the point process representation for the marked exceedance process (t_i, y_i). However, rather than treating this process as homogeneous in the time-domain with a constant mark distribution, we instead use a representation that allows for conditional nonstationarity in both domains. Specifically, we write the marked point process in terms of a conditional occurrence intensity and a conditional mark density,

$$\lambda(t, y \mid H_t) = \lambda(t \mid H_t) f(y \mid t, H_t),$$

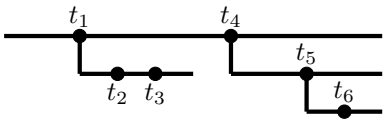
where H_t denotes the history of the process up to time t. This allows for both types of non-stationarity to be incorporated. First, the conditional intensity $\lambda(t \mid H_t)$ controls the times at which the extremes occur, and can incorporate features such as temporal clustering. Second, the conditional mark density $f(y \mid t, H_t)$ controls the magnitude of the extremes when they occur, and may also vary with the history of the process. We will model $\lambda(t \mid H_t)$ as a Hawkes process, which allows non-stationarity and clustered behaviour to arise directly from the conditional intensity function, and model $f(y \mid t, H_t)$ using a hierarchical GPD-based mark model which allows the distribution of exceedance magnitudes to vary across clusters.

3.1 Hawkes Process

A point process on the interval $[0, T]$ can be characterised by its conditional intensity function $\lambda(t \mid H_t)$, where H_t is the history of the process up to time t. This conditional



(a) Intensity function of a Hawkes process with events shown in Figure 2b



(b) Graphical representation of the structure of events arising in a simulated Hawkes process. Each black circle represents an event and shows how parent and child events form.

Fig. 2: Illustration of the branching structure and intensity function of a Hawkes process.

intensity function satisfies the counting property of a point process:

$$\lambda(t \mid H_t) = \lim_{\Delta t \to 0} \frac{\Pr(N(t+\Delta t) - N(t) = 1 \mid H_t)}{\Delta t},$$

where $N(t)$ is the number of events that occur in the interval $[0, t]$ (Daley and Vere-Jones 2003). The conditional intensity function is dependent on the history of the process so that past events can influence the rate at which future events occur. The Hawkes process (Hawkes 1971; Hawkes and Oakes 1974) is a type of self-exciting point process, with conditional intensity:

$$\lambda(t \mid H_t) = \mu(t) + \sum_{t_i < t} \kappa h(t - t_i), \tag{1}$$

where t_i denotes the time at which the ith event occurred, $\mu(t) > 0$ is a background intensity function, $\kappa > 0$ is a constant and $h(\cdot)$ is a probability density that integrates to 1, known as the triggering or excitation kernel. For ease of exposition we assume $\mu(t) = \mu$ is constant although time-variation in μ can be accounted for using a similar approach to (Deutsch and Ross 2025). The essence of the Hawkes process is that the occurrence of an event at time t makes it more likely for further events to occur soon afterwards, since the intensity increases by an amount controlled by the excitation kernel $\kappa h(\cdot)$. This leads to events occurring in clusters, making the Hawkes process well suited to modelling non-stationary exceedances. As an illustration, Figure 2a plots the conditional intensity function of a sample Hawkes process, where it can be seen

that each event increases the conditional intensity, leading to further events occurring and thus events arriving in clusters.

The Hawkes process can also be interpreted as a branching process as first noted by (Hawkes and Oakes 1974). At each time point t, suppose that n_t previous events have occurred. Then, the intensity function in Equation 1 can be interpreted as a linear superposition of $n_t + 1$ Poisson processes where the first has intensity function $\mu(t)$ and the other n_t triggered processes each have intensity function $\kappa h(t - t_j)$ for each previous event t_j. Under this interpretation, an event which occurs at time t_i will either have been generated by the background process $\mu(t)$ or by a triggered process, in which case we say that t_i is a child of the event t_j which triggered it.

Figure 2b shows a sample realisation of the Hawkes process, to illustrate how a branching structure appears naturally. Events t_1 and t_4 are exogenous (i.e. have no parent) and were generated by the background intensity $\mu(t)$. Events t_2, t_3, t_5 were all triggered by events from the background process, whereas event t_6 was triggered by t_5. This leads us to introduce a latent variable for each event $\mathbf{B} = \{B_1, B_2, \ldots, B_n\}$ that describes the branching structure, where B_i is the index of the event which caused t_i, with $B_i = 0$ if t_i is caused by the background process. Using Figure 2b as an example, the branching structure is: $\mathbf{B} = \{0, 1, 1, 0, 4, 5\}$. This branching variable and its effect on the posterior distribution leads to an efficient computational method for sampling the parameters of the Hawkes process which we will discuss later.

3.1.1 A Nonparametric Kernel

In most previous applications of the Hawkes process, the triggering kernel $h(t)$ has been specified parametrically and fitted using frequentist techniques (Porter and White 2012; Balderama et al. 2012). However in most realistic applications it will not be obvious which parametric form is most appropriate. As such, there has been recent interest in nonparametric specifications of the kernel function, typically within a frequentist framework using kernel density estimation or histogram estimators (Zhuang et al. 2002; Fox et al. 2016; Bacry et al. 2012). Additionally, Markwick, D and Ross, G. J. (2020) presented a Bayesian version of the Hawkes process which uses a Dirichlet process mixture prior to nonparametrically model the background rate $\mu(t)$ within a Bayesian setting. We extend their work to form a nonparametric estimate of $h(t)$.

The Dirichlet process (DP) is a stochastic process whose draws are distributions themselves, which is commonly used as a prior over the space of probability distributions. A DP is specified by its base distribution G_0 and a concentration parameter α_{DP}. If G is a draw from a DP then we write $G \sim DP(\alpha_{DP}, G_0)$. It can be shown (Ishwaran and James 2001; Sethuraman 1994) that G can be written in stick-breaking form as:

$$\begin{aligned} G &= \sum_{i=1}^{\infty} w_i \delta_{\psi_i}, \quad \psi_i \sim G_0, \\ w_i &= z_i \prod_{j=1}^{k-1} (1 - z_j), \quad z_j \sim \text{Beta}(1, \alpha_{DP}), \end{aligned} \tag{2}$$

where δ_{ψ_i} are point masses located at ψ_i and the z_j's are independent. From this, it can be seen that distributions drawn from a DP are almost surely discrete. The DP is a conjugate prior in the following sense: if $G \sim DP(\alpha_{DP}, G_0)$ and $\mathbf{z} = (z_1, \ldots, z_n)$ are independent draws from G then the posterior distribution for G is also a DP $p(G|\mathbf{z}) = DP(\alpha', G_0')$ where (Ferguson 1973; Antoniak 1974):

$$\begin{aligned} \alpha' &= \alpha + n \\ G_0' &= \frac{\alpha_{DP}}{\alpha_{DP} + n} G_0 + \frac{n}{\alpha_{DP} + n} \sum_{i=1}^{n} \frac{1}{n} \delta_{z_i}. \end{aligned} \tag{3}$$

Since our aim is to use the DP as a nonparametric prior on the excitation kernel $h(\cdot)$ of a continuous time process, the discreteness of DP samples is problematic. We hence follow a standard approach (Escobar and West 1995; Neal 2000) and convolve G with a continuous mixture kernel k. This produces a mixture model with an infinite number of components, leading to the specification:

$$\begin{aligned} h(t) &= \int k(t \mid \phi) \mathrm{d}G(\phi), \\ \phi &\sim G, \\ G &\sim \mathrm{DP}(\alpha_{DP}, G_0), \end{aligned} \tag{4}$$

where G_0 is the base distribution of the Dirichlet process and α_{DP} is the concentration parameter. By using an infinite mixture, the shape of $h(t)$ can be flexibly learned from the data in order to incorporate features such as heavy tails and multimodality that might be missed when using a parametric kernel. To aid computation, we choose a Lognormal distribution for the kernel $k(\cdot)$ so that $\phi = \{\mu, \sigma^2\}$. The Lognormal mixture model is a flexible choice which allows for potential multimodality and heavy tails in $h(\cdot)$. It also admits a conjugate prior which will simplify parameter inference. Specifically, when using the Lognormal kernel we choose the base measure to be $G_0 = N(\mu \mid \mu_0, \frac{\sigma^2}{k_0})\text{Inv-Gamma}(\sigma^2 \mid \alpha_0, \beta_0)$ where the prior parameters are taken to be weakly informative: $\{\mu_0 = 0, k_0 = 1, \alpha_0 = 1, \beta_0 = 1\}$.

3.2 Extreme Value Theory and the GPD

The observations in an extreme value event process consist of pairs of values (t_i, y_i), where t_i is the time at which the ith extreme event occurs and $y_i = r_i - u$ is the excess magnitude above a given threshold u. We use the Hawkes process to model the occurrence times of the extreme events, i.e. the t_i values. We now discuss our non-stationary model for the magnitudes y_i. In the simplest application of the PBH theorem we could assume that, for some threshold u, the observations are independent and identically distributed samples from a GPD,

$$y_i \sim \mathrm{GPD}(\sigma, \xi),$$

with common parameters σ and ξ. These parameters can then be estimated, which allows us to compute the future probability of large excesses. However, this assumes that the distribution of the y_i values is constant over time, which is not the case in many real-world applications. In the traditional extreme value literature these problems are often addressed by allowing the parameters of the GPD to depend on time, and applying common regression estimation techniques (Coles 2001). However, this requires choosing a particular parametric form for the time dependence, which can lead to under-fitting or over-fitting depending on the number of parameters chosen. It also does not easily allow for the formation of clusters in the data, which were clearly seen in Figure 1. Instead, we propose a method that groups the data into local temporal regimes using the branching structure of the Hawkes process, and allows the mark distribution to vary across these regimes. A hierarchical model is then used to pool information across clusters so that the cluster-level quantities can be estimated accurately even when the number of observations in each cluster is small.

Recall from Section 3.1 that the Hawkes process naturally produces clusters of events, represented by the branching variables $\mathbf{B} = (B_1, \ldots, B_n)$, where $B_i = j$ if (t_i, y_i) was generated by the Poisson process spawned by the event that occurred at time t_j, and $B_i = 0$ if (t_i, y_i) was generated by the background process. The occurrence of events in the background process naturally splits the observations into clusters, since each background event can result in a cascade of further events, as shown in Figure 2b. A natural alternative would be to define clusters genealogically, by grouping each background event with all of its descendants. However, when several background events occur close together, their descendants may be interleaved in time, which makes the resulting mark regimes less natural as local temporal regimes.

As such, we instead define each cluster as the temporal interval which elapses between each pair of background events. Let $a_1 < a_2 < \cdots < a_K$ denote the indices of the background events, so that $B_{a_k} = 0$ for $k = 1, \ldots, K$. Define $s_k = t_{a_k}$ for $k = 1, \ldots, K$, and set $s_{K+1} = T$. The kth cluster is then

$$C_k = \{y_i : s_k \leq t_i < s_{k+1}\}, \qquad k = 1, \ldots, K.$$

Since the first event is necessarily a background event, these sets form a partition of the observed exceedances. Referring back to Figure 2b, this sequence would hence have two clusters, the first consisting of events $\{t_1, t_2, t_3\}$ and the second consisting of events $\{t_4, t_5, t_6\}$.

A fully separate GPD fit within each cluster would be unstable, since the Hawkes branching structure can produce many clusters containing only a small number of exceedances. We therefore use the clustering to model local variation in the scale of exceedances, while estimating the tail-shape parameter globally. Conditional on the branching structure $\mathbf{B}$, if $y_i \in C_k$ we write

$$\begin{aligned} y_i \mid \mathbf{B}, \sigma_k, \xi &\sim \mathrm{GPD}(\sigma_k, \xi), \\ \log \sigma_k &= \log \sigma_0 + \tau_\sigma z_k, \qquad z_k \sim N(0,1), \end{aligned} \tag{5}$$

where σ_k is the GPD scale in cluster C_k and ξ is a shared shape parameter. We note that this specification is conditional on the branching structure $\mathbf{B}$ which divides the

sequence of observations into clusters. Since $\mathbf{B}$ is unknown, it must be estimated, which will be discussed in the next section. The marginal distribution of each y_i can then be obtained by marginalising $\mathbf{B}$ out of the joint posterior distribution for all model parameters.

The hierarchical prior on the cluster scales allows information to be borrowed across clusters (Gelman 2006; Gelman et al. 2013). We use weakly informative priors on the scale hierarchy and a regularising prior on the shared shape parameter:

$$\begin{aligned} \log \sigma_0 &\sim N(0, 1), \\ \tau_\sigma &\sim N^+(0, 0.5^2), \\ \xi &\sim N(0, 0.2^2) \quad \text{truncated to } (-0.25, \infty), \end{aligned} \tag{6}$$

where N^+ denotes a Normal distribution truncated to the positive real line. The shared-shape specification is a deliberate regularisation: clusters may differ in exceedance scale, which captures changes in the typical magnitude of exceedances, while the tail-shape parameter is learned globally from all exceedances. The lower truncation on ξ keeps inference away from the irregular finite-endpoint boundary of the GPD likelihood.

4 Posterior Inference

Posterior inference for our model involves estimating the Hawkes process parameters $\Theta_{HP} = \{\mu, \kappa, h(\cdot), \mathbf{B}\}$, which govern the exceedance times $t_1, \ldots, t_n$, and the GPD mark parameters

$$\Theta_{GPD} = \{z_1, \ldots, z_K, \log \sigma_0, \tau_\sigma, \xi\},$$

which govern the excess magnitudes y_i through $\sigma_k = \exp(\log \sigma_0 + \tau_\sigma z_k)$. Here K is not fixed in advance, but is determined by the current branching structure $\mathbf{B}$.

Our MCMC sampler uses a modular posterior factorisation in which the Hawkes process is learned from the exceedance times and the mark model is fitted conditional on posterior draws of the induced temporal clustering. This deliberately prevents the exceedance magnitudes from feeding back into the estimation of the temporal branching structure. The choice is motivated by the interpretation of the Hawkes process as a model for exceedance occurrence times, with the mark model used to describe variation in excess magnitudes across the resulting temporal regimes. We write $Y = \{(t_1, y_1), \ldots, (t_n, y_n)\}$ for the observed exceedances, and discuss each block of the sampler in turn.

4.1 Sampling for the Hawkes Process

To sample the Hawkes process parameters we use a similar approach to (Ross 2021; Rasmussen 2013), with an extension for the nonparametric Dirichlet process model for $h(\cdot)$. The latent branching variables $\mathbf{B} = (B_1, \ldots, B_n)$ decompose the Hawkes likelihood into approximately independent parameter blocks. Given this branching

structure, the event indices can be partitioned into the sets $S_0, S_1, \ldots, S_n$, where

$$S_j = \{i : B_i = j\}.$$

Thus S_0 is the set of all events that were created by the background process with intensity μ, and each S_j is the set of events that were produced by the process triggered by event t_j. Conditional on $\mathbf{B}$, the Hawkes likelihood for the event times can be written as (Rasmussen 2013)

$$L_{HP}(\mu, \kappa, h; t, \mathbf{B}) = \mu^{|S_0|} e^{-\mu T} \prod_{j=1}^{n} \left\{ e^{-\kappa H(T-t_j)} \kappa^{|S_j|} \prod_{i \in S_j} h(t_i - t_j) \right\}, \tag{7}$$

where $t = (t_1, \ldots, t_n)$, $H(z) = \int_0^z h(u)\, du$, and $|S_j|$ denotes the number of events in set S_j. The point of this latent variable parameterisation is that the posterior for μ is independent of the other Hawkes parameters, while the posterior dependence between κ and $h(\cdot)$ is substantially reduced. The MCMC sampling for $\mathbf{B}$, μ and κ then follows the same general scheme as in Ross (2021), before updating the nonparametric triggering kernel $h(\cdot)$.

Sampling from $p(\mathbf{B} \mid \Theta_{HP}^{(s-1)}, Y)$***:***

At each time point t, the Hawkes intensity in Equation (1) is a superposition of a background Poisson process with intensity μ and multiple triggered Poisson processes, one for each previous event. Conditional on the Hawkes parameters, the posterior allocation probabilities are proportional to the contribution of each component to the total intensity. Thus, for event t_i,

$$\begin{aligned} \Pr(B_i = 0 \mid \Theta_{HP}^{(s-1)}, Y) &= \frac{\mu}{\lambda(t_i \mid H_{t_i})}, \\ \Pr(B_i = j \mid \Theta_{HP}^{(s-1)}, Y) &= \frac{\kappa h(t_i - t_j)}{\lambda(t_i \mid H_{t_i})}, \qquad j = 1, \ldots, i-1, \end{aligned} \tag{8}$$

where $\lambda(t_i \mid H_{t_i}) = \mu + \sum_{j<i} \kappa h(t_i - t_j)$ is the total conditional intensity at t_i. Since these distributions are discrete, each B_i can be directly sampled from its posterior using multinomial sampling. The first event is necessarily assigned to the background process.

Sampling from $p(\mu \mid \mathbf{B}, Y)$***:***

Conditional on $\mathbf{B}$, the events in set S_0 follow a homogeneous Poisson process with intensity function μ. We use a conjugate prior $p(\mu) = \text{Gamma}(\alpha_\mu, \beta_\mu)$, where β_μ is a rate parameter, leading to the posterior distribution

$$\mu \mid \mathbf{B}, Y \sim \text{Gamma}\left(\alpha_\mu + |S_0|, \beta_\mu + T\right),$$

which can be sampled from exactly.

Sampling from $p(\kappa \mid \mathbf{B}, h, Y)$***:***

From Equation (7), the conditional posterior for κ is

$$p(\kappa \mid \mathbf{B}, h, Y) \propto p(\kappa)\kappa^{\sum_{j=1}^{n}|S_j|} \exp\left\{-\kappa \sum_{j=1}^{n} H(T - t_j)\right\}.$$

If $p(\kappa) = \text{Gamma}(\alpha_\kappa, \beta_\kappa)$, this gives the conjugate update

$$\kappa \mid \mathbf{B}, h, Y \sim \text{Gamma}\left(\alpha_\kappa + \sum_{j=1}^{n}|S_j|, \beta_\kappa + \sum_{j=1}^{n} H(T - t_j)\right),$$

again using the rate parametrisation. If the subcriticality condition $\kappa < 1$ is enforced through the prior, the same update is used with this Gamma distribution truncated to the interval $(0, 1)$.

Sampling from $p(h(\cdot) \mid \mathbf{B}, \mu, \kappa, Y)$***:***

For each event t_i that is not in the background process, let $x_i = t_i - t_{B_i}$ be the observed triggering lag. The likelihood contribution involving $h(\cdot)$ is

$$\left\{\prod_{i:B_i>0} h(x_i)\right\} \exp\left\{-\kappa \sum_{j=1}^{n} H(T - t_j)\right\}.$$

The first term is the usual density contribution from the observed triggering lags, while the second term is the integrated hazard contribution from the finite observation window. We update $h(\cdot)$ using the conjugate Dirichlet process mixture update for the observed lags, with a Metropolis–Hastings correction for the integrated hazard term.

The mixture model in Equation (4) can be rewritten as

$$x_i \sim k(x_i \mid \phi_i), \qquad \phi_i \sim G, \qquad G \sim \text{DP}(\alpha_{DP}, G_0).$$

We use the Chinese Restaurant Process sampler of Neal (2000, Algorithm 4) to update each mixture parameter ϕ_i given the current triggering lags. Given these ϕ_i values, the posterior for G is $p(G \mid \alpha_{DP}, G_0, \{\phi_i\})$, which from the conjugacy property in Equation (3) is also a DP distribution. A proposal $G^\star$ can then be drawn from this posterior using truncated stick breaking (Ishwaran and James 2001). Specifically, for a large truncation level L,

$$G^\star = \sum_{\ell=1}^{L} w_\ell^\star \delta_{\psi_\ell^\star}, \qquad \psi_\ell^\star \sim G_0',$$

with weights

$$w_\ell^\star = v_\ell^\star \prod_{r<\ell}(1 - v_r^\star), \qquad v_\ell^\star \sim \text{Beta}(1, \alpha'_{DP}),$$

where G'_0 and α'_{DP} are the posterior values from Equation (3). In our implementation we used $L = 1000$. Given $G^\star$, the proposed triggering kernel $h^\star$ and its cumulative distribution function $H^\star$ are fully defined, with

$$H^\star(z) = \sum_{\ell=1}^{L} w_\ell^\star \Phi(z \mid \psi_\ell^\star),$$

where Φ is the cumulative distribution function of the Lognormal mixture kernel k. Since this proposal is based on the posterior for the observed triggering lags, the Metropolis–Hastings acceptance probability only needs to correct for the integrated hazard term, giving

$$\min\left\{1, \exp\left[-\kappa \sum_{j=1}^{n} \{H^\star(T - t_j) - H(T - t_j)\}\right]\right\}.$$

4.2 Sampling for the GPD

Given a branching structure $\mathbf{B}$, the excess magnitudes $y_1, \ldots, y_n$ are divided into clusters based on Equation (5). Suppose there are K such clusters under the current branching structure. Conditional on this partition, the posterior for the mark model is proportional to

$$p(\log \sigma_0, \tau_\sigma, \xi) \prod_{k=1}^{K} \left\{ p(z_k) \prod_{i:y_i \in C_k} g(y_i \mid \sigma_k, \xi) \right\},$$

where $g(\cdot \mid \sigma_k, \xi)$ is the GPD density and

$$\sigma_k = \exp(\log \sigma_0 + \tau_\sigma z_k).$$

The GPD density is evaluated subject to its usual support constraint $1 + \xi y_i/\sigma_k > 0$. Conditional on the current clustering, we update the mark-model parameters using Hamiltonian Monte Carlo as implemented in Stan (Carpenter et al. 2017). Since the branching structure, and hence the clusters, can change between MCMC iterations, the cluster-specific scales are treated as local to the current partition rather than being assigned persistent labels across all MCMC iterations. This avoids the need to match cluster labels between different branching structures.

4.3 Posterior prediction

The fitted model can be used to obtain posterior predictive distributions for future exceedances. For each posterior draw of the model parameters, we simulate the Hawkes

process forward over a future window $(T, T+H]$ conditional on the observed history up to time T. This produces a draw of the future exceedance count

$$N_H = N(T+H) - N(T),$$

along with the corresponding future exceedance times. Conditional on the simulated branching structure, excess magnitudes are then drawn from the appropriate GPD distribution. Events assigned to an existing cluster use the corresponding cluster-level scale parameter and the shared shape parameter. New background events initiate new clusters; for these clusters a new latent $z_{\text{new}} \sim N(0,1)$ is drawn, giving

$$\log \sigma_{\text{new}} = \log \sigma_0 + \tau_\sigma z_{\text{new}},$$

while the same global shape parameter ξ is used.

Repeating this procedure across posterior draws gives predictive distributions for quantities such as the number of exceedances in the future window, the maximum future excess

$$M_H = \max\{Y_i : T < t_i \leq T+H\},$$

and tail probabilities of the form

$$\Pr(M_H > z \mid Y),$$

for high levels z. These are the predictive quantities used in the empirical analysis below, in addition to the out-of-sample predictive likelihood. For numerical stability the GPD model is fitted to excesses divided by a training-set scale factor. Predictive densities reported on the original excess scale include the corresponding Jacobian correction.

4.4 Computational considerations

The most expensive part of the sampler is the update of the branching variables, since each event can in principle be assigned to any earlier event. A direct implementation of this step is therefore $O(n^2)$ in the number of exceedances. In the present setting this is typically manageable because the model is fitted only to threshold exceedances, so n is much smaller than the length of the original time series. In larger applications the branching update can be accelerated by ignoring parent assignments whose triggering lag has negligible density under the current kernel. The truncation level L in the stick-breaking representation is chosen sufficiently large that the remaining stick mass is negligible. In practice we monitor convergence using trace plots and posterior summaries for μ, κ, the GPD scale-hierarchy parameters, and the main posterior predictive quantities.

Table 1 summarises the prior distributions and fixed hyperparameter settings used in the fitted models. For each real-data analysis we ran four chains for both the Exponential-Hawkes and DP-Hawkes models, using 10,000 iterations per chain and discarding the first 2,000 iterations as burn-in. No thinning was used. The dirichletprocess R package (Ross and Markwick 2018) was used to implement the above sampling

Component	Parameter	Prior or setting
Hawkes background	μ	Gamma$(0.1, 0.1)$
Hawkes branching	κ	Uniform on $(0, 1)$
Exponential kernel	β	Uniform on $(0, 100)$
DP concentration	α_{DP}	Gamma$(2, 4)$
DP log-lag base measure	G_0	Gaussian base measure with parameters $(0, 1, 1, 1)$
GPD global log-scale	$\log \sigma_0$	$N(0, 1)$
GPD scale variation	τ_σ	half-$N(0, 0.5^2)$
GPD shape	ξ	$N(0, 0.2^2)$, truncated below at -0.25

Table 1: Prior distributions and fixed hyperparameter settings used in the fitted models. Gamma distributions use the shape/rate parameterisation. The DP mixture is fitted to log triggering lags; the Gaussian base-measure parameters are $(0, 1, 1, 1)$, and the DP concentration parameter uses the Gamma$(2, 4)$ prior.

of the DP. For the hierarchical GPD mark model, we used 100 evenly spaced retained Hawkes posterior draws per model as representative branchings. The final hierarchical GPD fits were run in Stan using four chains, 2,000 iterations per chain, 1,000 warm-up iterations, `adapt_delta`=0.99, and `max_treedepth`=15.

5 Simulation Study

We first conduct a simulation study to assess our models in a setting where the true data-generating mechanism is known. We consider a 2×2 design in which the true Hawkes triggering kernel is either Exponential or a two-component mixture, and the true mark distribution is either an iid GPD or a hierarchical GPD with clusters.

In all scenarios, events are generated on the interval $[0, 1000]$ from a Hawkes process with background rate $\mu = 0.10$ and branching parameter $\kappa = 0.55$. The first 800 time units are used for training and the remaining 200 for testing and prediction. In the exponential-kernel scenarios the triggering density is $h(t) = \beta \exp(-\beta t)$ with $\beta = 1$. In the mixture-kernel scenarios the triggering density is

$$h(t) = 0.7\,\text{Lognormal}(-0.3, 0.35^2) + 0.3\,\text{Lognormal}(1.2, 0.45^2),$$

which produces a non-exponential excitation pattern with both short- and longer-lag triggering behaviour. For the mark distribution we set $\sigma_0 = 1$ and $\xi = 0.15$. In the iid mark scenarios, all excesses are generated from a common GPD(σ_0, ξ) distribution. In the hierarchical mark scenarios, excesses in cluster C_k are generated from GPD(σ_k, ξ), where

$$\log \sigma_k = \log \sigma_0 + \tau_\sigma z_k, \qquad z_k \sim N(0, 1),$$

with $\tau_\sigma = 1$. As in the fitted model, clusters are defined as temporal intervals between background events in the Hawkes branching structure.

For each simulated data set we fit the following four models:

1. Hawkes process with a parametric Exponential kernel $h(t)$, with the y_t values treated as independent draws from a single GPD(σ, ξ) distribution.

Truth scenario	Exp+iid	DP+iid	Exp+hier.	DP+hier.
Exponential kernel, iid marks	0.000	0.035 (0.176)	0.104 (0.090)	0.176 (0.159)
Exponential kernel, hier. marks	0.000	-0.176 (0.197)	2.125 (0.935)	2.322 (1.323)
Mixture kernel, iid marks	0.000	1.703 (0.619)	-0.075 (0.061)	1.665 (0.548)
Mixture kernel, hier. marks	0.000	1.922 (0.445)	0.806 (0.689)	2.675 (0.871)

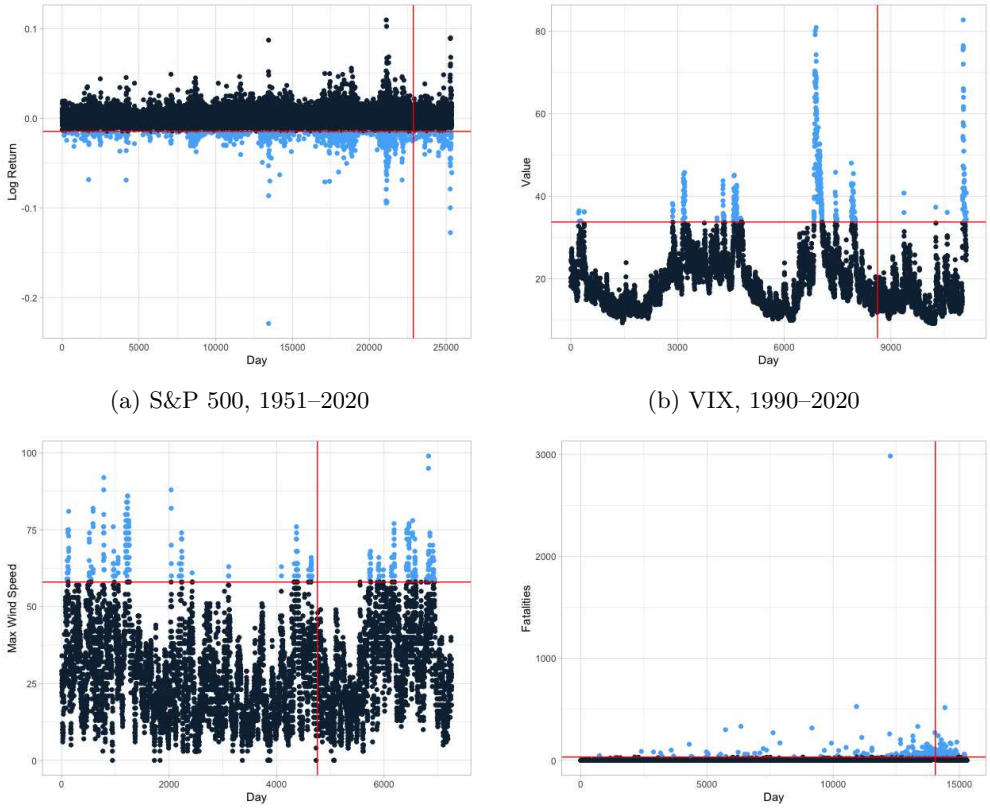
Table 2: Mean held-out (test set) log predictive score differences in the simulation study, relative to the Exp+iid model within each replicate. Monte Carlo standard errors are shown in parentheses. Positive values indicate improvement over the Exp+iid baseline.

2. Hawkes process with a nonparametric Lognormal DP mixture kernel $h(t)$, with the y_t values treated as independent draws from a single GPD(σ, ξ) distribution.
3. Hawkes process with a parametric Exponential kernel $h(t)$, with the y_t values modelled using the hierarchical cluster-scale GPD model described in Section 3.
4. Hawkes process with a nonparametric Lognormal DP mixture kernel $h(t)$, with the y_t values modelled using the hierarchical cluster-scale GPD model described in Section 3.

Performance is measured using held-out log predictive scores on the test period. To make the comparisons stable across simulated data sets, Table 2 reports score differences relative to the Exp+iid baseline (Model 1 above) within each replicate. Positive values therefore indicate an improvement over the baseline model. We ran 10 replicates for each scenario.

The results in Table 2 show the intended behaviour of the four model components. When the data are generated from an exponential Hawkes process with iid GPD marks, the more flexible models are essentially tied with the correctly specified Exp+iid baseline, indicating that they incur little penalty from their additional flexibility when it is not required. When cluster-level mark variation is introduced while keeping the exponential triggering kernel, the hierarchical GPD models improve the held-out predictive score. Conversely, when the Hawkes triggering kernel is generated from the non-exponential mixture while the marks remain iid, the DP-Hawkes models improve over the exponential-kernel models, while the hierarchical mark component provides little additional benefit. Finally, when both the mixture triggering kernel and hierarchical mark variation are present, the full DP-Hawkes plus hierarchical GPD model gives the largest improvement.

In summary, the simulation study shows that the two flexible components of the model behave as intended. The DP-Hawkes kernel improves prediction when the true triggering mechanism is non-exponential, while the hierarchical GPD mark model improves prediction when exceedance magnitudes vary across Hawkes-induced clusters. When the corresponding feature is absent from the data-generating mechanism, the additional flexibility incurs little predictive cost.



(a) S&P 500, 1951–2020

(b) VIX, 1990–2020

(c) Bradfield wind speeds, 1975–1984

(d) RAND-MIPT Terrorism, 1968–2009

Fig. 3: Plots of the four real data sets being considered. Light blue points denote the extremes, with the horizontal red line denoting the threshold u. The vertical red line shows the separation into training and test sets.

6 Real Data Applications

We next evaluate the model on four real data sets from finance, environmental extremes, and terrorism. The data sets are shown in Figure 3. In each case the original time series is converted into a sequence of threshold exceedances (t_i, y_i), where t_i is the exceedance time and $y_i = r_i - u$ is the excess above the threshold.

S&P 500:

The S&P 500 is a stock market index composed of 500 large companies listed on American stock market exchanges, and is generally thought to be a good representation of the overall U.S. market. Let x_t denote the closing price of the index on day t. Since interest is in extreme losses, we define the marks using negative log-returns. The data set consists of daily observations from 1951 to 2020.

VIX:

The VIX (CBOE Volatility Index) is derived from S&P 500 options and provides a measure of expected market volatility. We obtained daily closing prices of the VIX from 1990 to 2020. Unlike the S&P 500, the VIX is approximately mean-stationary over a long time horizon, so we define the marks r_t to be the closing price on day t rather than first differences.

Bradfield wind speed:

Extreme wind-speed analysis is a standard application area for Extreme Value Theory (Palutikof et al. 1999). We obtained[1] a time series consisting of the maximum hourly observed gust wind speed, in knots, in High Bradfield in the United Kingdom between 1975 and 1984. Here r_t denotes the maximum wind speed during hour t.

Terrorism:

The RAND Database of Worldwide Terrorism Incidents (RDWTI) is publicly available[2] and contains a detailed list of global terrorism events that occurred between 1968 and 2009. This data set has previously been studied in an extreme-value context by Porter and White (2012). For each day during the sample period, we aggregated the total number of worldwide fatalities, so that r_t denotes the number of fatalities on day t. The extremely large outlier in Figure 3 corresponds to the 9/11 World Trade Centre attack, while the subsequent increase in terrorism-related fatalities is mostly due to the insurgency in Iraq following the resulting invasion.

6.1 Evaluation protocol

For each data set, we choose the extreme-value threshold u to be the 95th percentile of the observed r_t values. In the case of the S&P 500, where interest lies in extreme losses rather than extreme gains, we use the lower 5th percentile instead. More advanced threshold-selection methods are available (Coles 2001), but a fixed percentile gives a simple and comparable evaluation across the four applications.

Our proposed model has two main components beyond a standard Hawkes-POT construction: a nonparametric Hawkes triggering kernel and a hierarchical cluster-scale GPD mark model. We therefore compare the same four models as in the simulation study: Exponential or DP-Hawkes kernels, crossed with either a common iid GPD mark distribution or the hierarchical cluster-scale GPD mark model.

Performance is assessed using held-out predictive log scores. Each data set is divided into a training set Y and a test set $\tilde{Y}$; the models are estimated on the training set and used to predict both the times and magnitudes of the test-set exceedances. The Bayesian predictive likelihood is

$$p(\tilde{Y} \mid Y) = \int p(\tilde{Y} \mid \Theta) p(\Theta \mid Y) d\Theta \approx \frac{1}{M} \sum_{s=1}^{M} p(\tilde{Y} \mid \Theta^{(s)}),$$

[1] http://www.mas.ncl.ac.uk/~nlf8/shortcourse/part3.pdf
[2] https://www.rand.org/nsrd/projects/terrorism-incidents/download.html

Dataset	Exp+iid	DP+iid	Exp+hier.	DP+hier.
S&P 500	-46.62	-41.84	-36.42	**-33.47**
VIX	-312.25	-284.16	-301.03	**-280.30**
Bradfield	-730.14	-672.77	-714.76	**-659.00**
Terrorism	-1525.40	-1519.45	-1525.43	**-1519.18**

Table 3: Held-out combined log predictive scores on the original excess scale for the four real data sets. Higher is better; the best score in each row is bolded. Here Exp denotes the parametric Exponential Hawkes kernel, DP denotes the nonparametric DP mixture Hawkes kernel, iid denotes the common GPD mark model, and hier. denotes the hierarchical cluster-scale GPD mark model.

where $\Theta^{(s)}$ are posterior samples obtained from the training set. For the time component, the held-out point-process likelihood was evaluated sequentially over the observed test events. Each test event was scored conditional on the full training history and all earlier observed test events in the test window. Thus, after a test event occurs, it is allowed to affect the conditional intensity assigned to later test events, just as it would in the likelihood for an observed Hawkes process path. The likelihood also includes the probability of observing no additional events between successive observed test events and after the final test event in the test window. For all data sets except Terrorism, the most recent 10 years are used as the test set. For Terrorism, most extreme events occur in the final decade, so we instead use the most recent five years as the test set to retain sufficient training data for parameter estimation. For numerical stability, the GPD model was fitted to excesses divided by the median positive training excess, with the mean positive training excess used as a fallback if needed. Held-out mark log scores were transformed back to the original excess scale by including the corresponding Jacobian correction.

Figure 4 shows a representative posterior clustering under the DP-Hawkes model for each of the four data sets. These plots are intended as illustrations of the local temporal regimes induced by the Hawkes branching structure, rather than as unique cluster assignments. The induced clustering is then used by the mark model to allow the GPD scale to vary across regimes.

6.2 Predictive model comparison

Table 3 gives the held-out combined log predictive scores for the four model combinations. Higher values indicate better predictive performance. The nonparametric DP-Hawkes kernel improves on the Exponential kernel in every data set when paired with the hierarchical GPD mark model. The full DP-Hawkes plus hierarchical GPD model gives the best score in all four data sets, although the improvement over the DP-Hawkes plus iid GPD model is very small for Terrorism.

The comparison in Table 3 separates the two sources of improvement. Replacing the Exponential Hawkes kernel by the DP mixture kernel generally improves the prediction

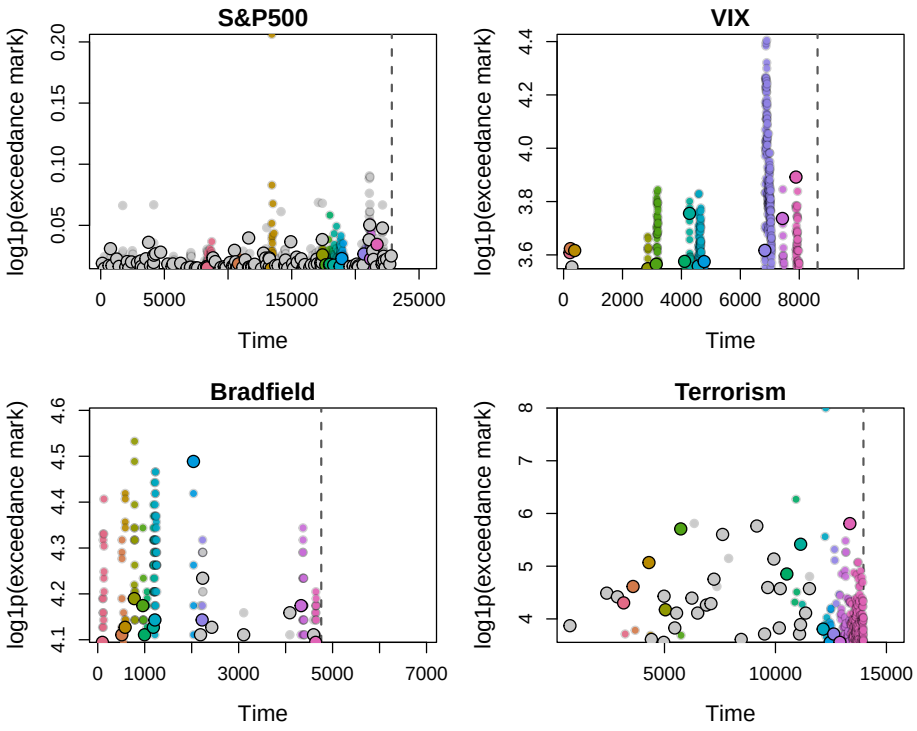


Fig. 4: Representative posterior DP-Hawkes clustering draws for the four data sets. Colours distinguish clusters within the displayed posterior draw only; cluster labels are not intended to be matched across MCMC draws. The vertical axis uses a $\log(1+y)$ scale for readability, and the vertical dashed line shows the split between training and test periods.

of exceedance times, while replacing the iid GPD mark model by the hierarchical cluster-scale GPD improves the prediction of exceedance magnitudes. The latter effect is isolated in Table 4, which compares the iid and hierarchical GPD mark models conditional on the DP-Hawkes time model. The hierarchical GPD improves the mark score in all four data sets, with substantial gains for S&P 500, VIX and Bradfield, and a small gain for Terrorism.

6.3 Posterior predictive behaviour and fitted mark parameters

A predictive likelihood can sometimes hide poor tail behaviour. We therefore also examine posterior predictive summaries for the maximum held-out excess under the full DP-Hawkes plus hierarchical GPD model. Table 5 compares the observed held-out maximum with the posterior predictive median and central 90% predictive interval. The observed maxima for S&P 500, VIX and Bradfield are broadly consistent with the posterior predictive distributions. The terrorism data set has a much heavier fitted

Dataset	iid GPD mark score	hierarchical GPD mark score	improvement
S&P 500	335.07	344.15	**9.08**
VIX	-192.16	-181.28	**10.87**
Bradfield	-417.75	-402.61	**15.14**
Terrorism	-981.68	-981.27	**0.41**

Table 4: DP-Hawkes mark-only comparison on the original excess scale. The improvement column is the hierarchical GPD score minus the iid GPD score, and isolates the contribution of the hierarchical mark model conditional on the DP-Hawkes time model. Higher is better.

Dataset	observed maximum	predictive median	90% predictive interval
S&P 500	0.113	0.045	(0.019, 0.124)
VIX	48.95	14.45	(1.52, 59.80)
Bradfield	41.00	34.27	(14.94, 71.05)
Terrorism	484.0	634.7	(194.0, 2464)

Table 5: Posterior predictive summaries for the maximum held-out excess under the full DP-Hawkes plus hierarchical GPD model. Predictive intervals are central 90% intervals, and all values are reported on the original excess scale.

tail, leading to a wider predictive distribution for the maximum future excess; this is consistent with the posterior estimate of the positive GPD shape parameter for that data set.

Figure 5 summarises the posterior distribution of the main GPD mark parameters under the full DP-Hawkes model. The global shape parameter ξ varies substantially between applications: it is positive for Terrorism, mildly positive for S&P 500, and close to the lower regularisation boundary for VIX and Bradfield. The posterior for τ_σ is away from zero in all four data sets, indicating that the Hawkes-induced clusters contain useful information about variation in the scale of exceedance magnitudes.

7 Discussion

In this paper we have developed and applied a novel framework for modelling extreme events that relaxes some of the conditions of standard extreme value theory and allows for nonstationarity in both the exceedance times and magnitudes. Our key focus is on the predictive modelling of future extreme values, which distinguishes our work from previous nonparametric Bayesian point process models which are essentially based on retrospective smoothing of historical extremes (Kottas and Sansó 2007; Kottas et al. 2012). We also developed a posterior simulation algorithm for the resulting model, and showed through both simulation and real-data examples that predictive accuracy can be improved by taking both types of nonstationarity into account.

The simulation and empirical results support both components of the proposed model. In the real-data applications, replacing the parametric Exponential Hawkes

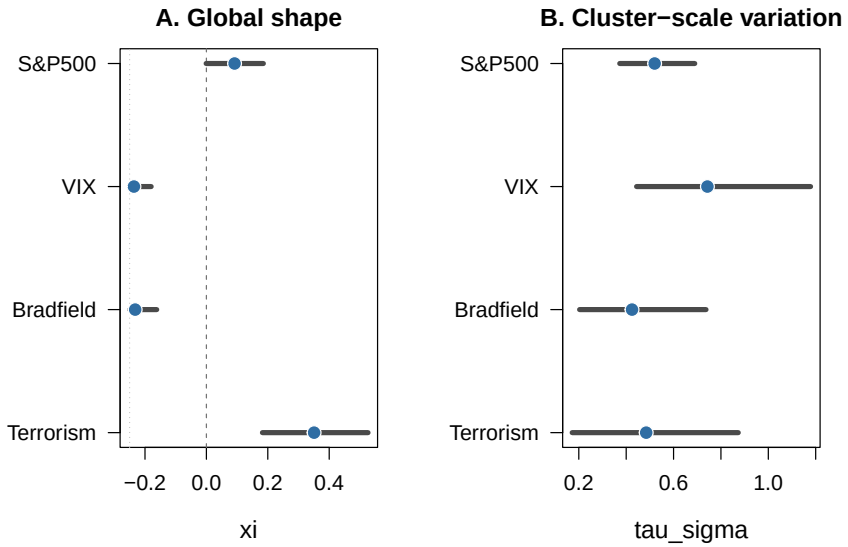


Fig. 5: Posterior medians and 95% credible intervals for the main GPD mark parameters under the full DP-Hawkes plus hierarchical GPD model. Panel A shows the shared GPD shape parameter ξ, with reference lines at $\xi = 0$ and at the lower truncation point $\xi = -0.25$. Panel B shows the cluster-scale variation parameter τ_σ.

kernel with a nonparametric Dirichlet process mixture kernel improved held-out predictive performance across the four data sets when paired with the hierarchical mark model. The simulation study shows the corresponding controlled behaviour: the DP kernel improves prediction when the true triggering pattern is non-exponential. This suggests that the temporal clustering of extreme events is not always adequately captured by a simple parametric excitation kernel. Second, conditional on the DP-Hawkes time model, replacing the iid GPD mark distribution with the hierarchical cluster-scale GPD improved the mark score in all four applications. The improvement was substantial for S&P 500, VIX and Bradfield wind speeds, and smaller for the terrorism data, where the iid and hierarchical mark models gave similar predictive performance. Taken together, these results indicate that Hawkes-induced clustering is useful not only for predicting when extremes occur, but also for modelling changes in their magnitudes. Together, the simulation and real-data results indicate that the two flexible components are useful for different aspects of nonstationarity.

A key modelling choice in the mark distribution is to allow the GPD scale to vary across Hawkes-induced clusters while estimating the shape parameter globally. This gives a compromise between a stationary POT model, which treats all exceedance magnitudes as iid draws from a single GPD, and a fully local model which attempts to estimate separate tail parameters within each cluster. The latter is unattractive in this setting because the branching structure can produce many small clusters, making cluster-specific tail-shape estimation unstable. The shared-shape specification retains

the ability to capture local changes in the typical magnitude of exceedances while borrowing information across the full data set for tail-shape inference. The posterior summaries in the real-data examples show that the fitted shape parameter varies substantially between applications, while the cluster-scale variation parameter is away from zero in each case.

Several extensions would be worth pursuing. The background intensity $\mu(t)$ could be made time-varying or allowed to depend on covariates, allowing long-run changes in the rate of exceedances to be separated from short-run self-excitation. Similarly, covariates could be introduced into the mark distribution, either through the cluster-scale hierarchy or through the global shape parameter. A multivariate extension would also be valuable in applications such as financial risk or natural hazards, where extremes in different series or regions may interact. More generally, the results suggest that combining self-exciting point process models with hierarchical extreme value models is a promising approach for forecasting nonstationary extremes.

Statements and Declarations

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was received for this work.

Data availability

The data sets analysed in this article are publicly available from the sources described in Section 6.

Code availability

Code implementing the methods and reproducing the analyses is available from the corresponding author upon reasonable request.

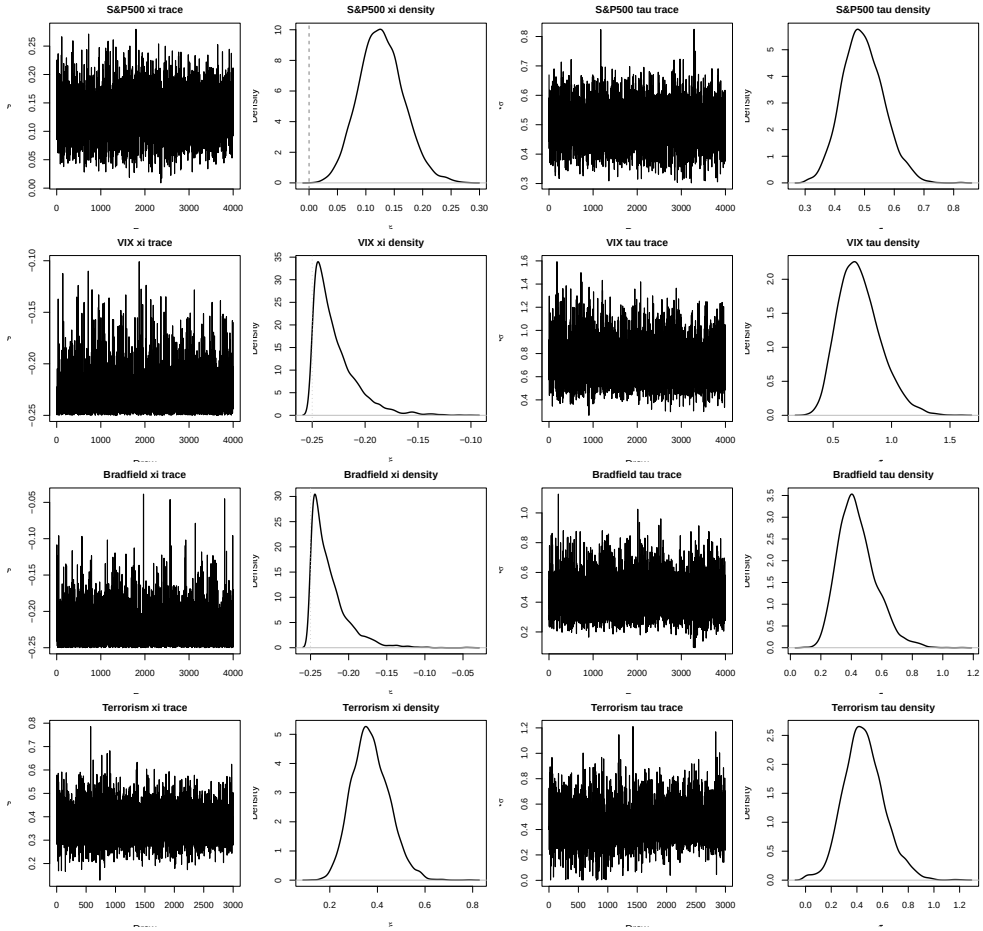


Fig. A.1: Trace and posterior density plots for ξ and τ_σ under the final DP-Hawkes plus hierarchical GPD model.

A MCMC diagnostics

Figure A.1 shows trace and posterior density plots for the global GPD shape parameter ξ and the cluster-scale variation parameter τ_σ under the final DP-Hawkes plus hierarchical GPD model. These diagnostics provide an additional check on the final real-data fits.

References

Porter, M., White, G.: Self-exciting hurdle models for terrorist activity. Annals of Applied Statistics **6**(1), 106–124 (2012)

Bray, A., Schoenberg, F.P.: Assessment of point process models for earthquake forecasting. Statistical science, 510–520 (2013)

Kiriliouk, A., Rootzén, H., Segers, J., Wadsworth, J.L.: Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions. Technometrics **61**(1) (2019)

Chavez-Demoulin, V., McGill, J.A.: High-frequency financial data modeling using Hawkes processes. Journal of Banking & Finance **36**(12), 3415–3426 (2012)

Coles, S.: An Introduction to Statistical Modeling of Extreme Values. Springer Series in Statistics. Springer, London (2001)

Davison, A.C., Smith, R.L.: Models for Exceedances over High Thresholds. Journal of the Royal Statistical Society. Series B (Methodological) **52**(3), 393–442 (1990)

Balkema, A.A., Haan, L.: Residual Life Time at Great Age. The Annals of Probability **2**(5), 792–804 (1974)

Chavez-Demoulin, V., Davison, A.C., McNeil, A.J.: Estimating value-at-risk: a point process approach. Quantitative Finance **5**(2), 227–234 (2005)

Poon, S.-H., Rockinger, M., Tawn, J.: Modelling Extreme-Value Dependence in International Stock Markets. Statistica Sinica **13**(4), 929–953 (2003)

Northrop, P.J., Jonathan, P.: Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. Environmetrics **22**(7), 799–809 (2011)

Gyarmati-Szabo, J.: Statistical extreme value modelling to study roadside air pollution episodes. Ph.D., University of Leeds (2011)

Stindl, T.: Forecasting intraday market risk: A marked self-exciting point process with exogenous renewals. Journal of Empirical Finance **70**, 182–198 (2023)

Eastoe, E.F., Tawn, J.A.: Modelling Non-Stationary Extremes with Application to Surface Level Ozone. Journal of the Royal Statistical Society. Series C (Applied Statistics) **58**(1), 25–45 (2009)

Kottas, A., Sansó, B.: Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis. Journal of Statistical Planning and Inference **137**(10), 3151–3163 (2007)

Kottas, A., Wang, Z., Rodríguez, A.: Spatial modeling for risk assessment of extreme values from environmental time series: a Bayesian nonparametric approach. Environmetrics **23**(8), 649–662 (2012)

Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)

Leadbetter, M.: Weak convergence of high level exceedances by a stationary sequence. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **34**(1), 11–15 (1976)

Hsing, T., Hüsler, J., Leadbetter, M.: On the exceedance point process for a stationary sequence. Probability Theory and Related Fields **78**(1), 97–112 (1988)

Daley, D., Vere-Jones, D.: An Introduction to the Theory of Point Processes. Volume I. Springer, New York (2003)

Hawkes, A.G., Oakes, D.: A Cluster Process Representation of a Self-Exciting Process. Journal of Applied Probability **11**(3), 493–503 (1974)

Deutsch, I., Ross, G.J.: Estimating product cannibalisation in wholesale using multivariate Hawkes processes with inhibition. The Annals of Applied Statistics **19**(1), 235–260 (2025)

Balderama, E., Schoenberg, F.P., Murray, E., Rundel, P.W.: Application of Branching Models in the Study of Invasive Species. Journal of the American Statistical Association **107**(498), 467–476 (2012)

Zhuang, J., Ogata, Y., Vere-Jones, D.: Stochastic Declustering of Space-Time Earthquake Occurrences. Journal of the American Statistical Association **97**(458), 369–380 (2002)

Fox, E.W., Schoenberg, F.P., Gordon, J.S.: Spatially inhomogeneous background rate estimators and uncertainty quantification for nonparametric Hawkes point process models of earthquake occurrences. Annals of Applied Statistics **10**(3), 1725–1756 (2016)

Bacry, E., Dayri, K., Muzy, J.F.: Non-parametric kernel estimation for symmetric Hawkes processes. Application to high frequency financial data. The European Physical Journal B **85**(5), 157 (2012)

Markwick, D and Ross, G. J.: Hierarchical nonparametric Hawkes process modelling of financial trading times. Preprint (2020)

Ishwaran, H., James, L.F.: Gibbs Sampling Methods for Stick-Breaking Priors. Journal of the American Statistical Association **96**(453), 161–173 (2001)

Sethuraman, J.: A Constructive Definition of Dricihlet Priors. Statistica Sinica **4**(2),

639–650 (1994)

Ferguson, T.S.: A Bayesian Analysis of Some Nonparametric Problems. Annals of Statistics **1**(2), 209–230 (1973)

Antoniak, C.E.: Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. Annals of Statistics **2**(6), 1152–1174 (1974)

Escobar, M.D., West, M.: Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association **90**(430), 577–588 (1995)

Neal, R.M.: Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Journal of Computational and Graphical Statistics **9**(2), 249–265 (2000)

Gelman, A.: Multilevel (Hierarchical) Modeling: What It Can and Cannot Do. Technometrics **48**(3), 432–435 (2006)

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, 3rd edn. CRC Press, Boca Raton, FL (2013)

Ross, G.J.: Bayesian estimation of the ETAS model for earthquake occurrences. Bulletin of the Seismological Society of America **111**(3), 1473–1480 (2021)

Rasmussen, J.G.: Bayesian Inference for Hawkes Processes. Methodology and Computing in Applied Probability **15**(15) (2013)

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A.: Stan: A Probabilistic Programming Language. Journal of Statistical Software **76**(1), 1–32 (2017)

Ross, G.J., Markwick, D.: dirichletprocess: An R Package for Fitting Complex Bayesian Nonparametric Models. R package vignette. Available from CRAN (2018)

Palutikof, J.P., Brabson, B.B., Lister, D.H., Adcock, S.T.: A review of methods to calculate extreme wind speeds. Meteorological Applications **6**(2), 119–132 (1999)